THE DELTA MULTIPLE COMPARISON METHOD.
PERFORMANCE AND USEFULNESS

Erling Sverdrup

1. Introduction

In his paper "On large-sample multiple comparison methods" (1988), Nils Lid Hjort has made some interesting comments and supplementary remarks to my paper "Multiple comparison and the likelihood ratio testing" (1986). This inspires me to comment upon his comments, thus expanding upon my discussion of the statistical ideas in my paper. I am grateful for having been given such an opportunity.

The content of the present paper is best summed up by giving its subtitles

1. Introduction.
2. Limitation of the generalized Scheffé multiple comparison method. The Bonferroni approach.
3. But the generalized Scheffé method works good in important situations. Comparison of performances. (See the important tables 1 and 2.)
4. The concept of performance function.
5. The large sample approach.
6. General and focalized parameters and contrasts.
7. The generality of the large sample multiple comparison method.
8. Do there exist multiple comparison methods not "hampered" by null-states?
9. Why multiple comparison tests? Why not just simultaneous confidence intervals?

2. Limitation of the generalized Scheffé multiple comparison method. The Bonferroni approach

I have felt that my description of the situation and the method given on p. 14-20 of my paper is complete in the sense that if the decision structure is a different one, then on the whole other methods would work better. Hjort's contribution has strengthened my intuition. This demotes the chi-square goodness of fit test, the Fisher's F-test and general likelihood ratio testing (with many parameters involved) from their traditionally high positions as handy all round methods.

Thus it is almost unanimously recommended to use chi-square testing of contingency tables, where, obviously, use of Bonferroni is more efficient; as nicely demonstrated by Hjort in section 4 (p.8-10) of his paper.

This attitude enhances the importance of the one-parametric tests, such as hypergeometric tests of categorical data and Student tests for normal data, perhaps in a complex context with stratification variables and regressors. We know that these exact tests have optimum properties in the Neyman-Pearson sense.

In the case of categorical data the exact one-parametric optimum tests have hitherto received little attention, although they obviously exist, even if the models are complex with many parameters.

As an example consider the situation with a 4-way classification with $4 \times 2 \times 3 \times 2$ levels and assume about the log-linear interactions that $\mu_{1234} = \mu_{123} = \mu_{124} = \mu_{134} = \mu_{234} = \mu_{13} = 0$ (for all combinations of levels (i,j,k,l)). We want to test if $\mu_{24}(2,2) > 0$ or < 0. Except for a randomization the Neyman-Pearson optimum test is the following. Let x_{ijkl} be the observed table entrances and t_{ijkl} hypothetical entrances, $N = \sum x_{ijkl} = \sum t_{ijkl}$.

Furthermore

$$S = (t \mid t_{ij++} = x_{ij++}, t_{i++l} = x_{i++l},\ t_{+jk+} = x_{+jk+}, t_{++kl} = x_{++kl},$$

for all $i,j,k,l,$),

$$S_v = S \cap (t \mid t_{+2+2} \leq v),\ K(v) = \sum_{t \in S_v} (\Pi t_{ijkl}!)^{-1}$$

Then $\mu_{24}(2,2)$ should be declared $<$ or $>$ 0 according as $K(V)/K(N)$ are $\leq \varepsilon$ or $> 1-\varepsilon$. Of course the fact that the test is exact makes it useful in cases where the observations are so few that most information is given by the pattern of zeros in the free-quency tables.

Systematization of such situations together with rules for constructing and programming the tests is important. (By the way, the importance of the Neyman-Pearson theory has been greatly enhanced by the modern computer technology.)

3. But the generalized Scheffé method works good in important situations. Comparison of performances

The Scheffé multiple comparison and its generalization (below called "the Scheffé method" for short), is really intended for situations where a priori there is an infinity of effects (contrasts) which may be of interest. The number of effects actually declared present may be finite or infinite. In neither case can the method based on Bonferroni's inequality be used.

However, it is tempting to use the Scheffé method in the case of a finite number of a priori feasible effects. Hjort shows by comparing the length of confidence intervals for single contrasts that in such cases "Bonferroni usually wins". This kind of comparison of methods, due to Scheffé (1958, p.76-77), indicates which method is the better one, but it does not tell us how much better in terms of probabilities. It does not give the performance function explicitly, i.e. the probabilities of discovering effects which are present. Thus, in terms of confidence intervals, for any parameter values η and η^0, and any effect f, it is of interest to study $\Pr(a(x) < f(\eta^0) < b(x)|\eta)$; not only $b(x) - a(x)$. With the computers available today it should be possible to study at least certain features of the performance function.

Let us see what this will bring forth in a particular example. Consider the one-way lay-out in analysis of variance with expectations $\xi_1,\ldots,\xi_r$ and variance σ^2. The numbers of observations in each of the r classes are $n_1, n_2,\ldots,n_r$, respectively and $n = \sum n_j$. We are interested in contrasts relatively to the null-state $\xi_1 = \xi_2 =\ldots= \xi_r$, i.e. in contrasts of the type $\sum_j f_j\xi_j$, where $\sum f_j = 0$. (We might have included non-linear contrasts, but that is unimportant for the present illustration.)

Under normality assumption this is a classical Scheffé situation. [We use the following notation for standard cumulative distribution functions. $G(x) = \Pr(X \leqslant x)$ where X is normal $(0,1)$. $\Gamma_m(z;\lambda) = \Pr(Z \leqslant z)$, where $Z = \sum_1^m X_j^2$ and $X_1,\ldots,X_m$ are independent norm $(\nu_j,1)$ $j=1,2,\ldots,m$ respectively, and $\lambda = \sum_1^m \nu_j^2$. $G_{m,n}(f;\lambda) = \Pr(F \leqslant f)$ where $F = Z_1 n/Z_2 m$, Z_1 and Z_2 are independent with cumulative distribution functions, respectively $\Gamma_m(z;\lambda)$ and $\Gamma_n(z;0)$. $G_n(t;\nu) = \Pr(T \leqslant t)$, where $T = X\sqrt{n}/\sqrt{Z}$, X and Z are independent, X is normal $(\nu,1)$ and Z has distribution $\Gamma_n(z;0)$. - We use lower case letters for the corresponding densities. We write $\Gamma_\nu(z)$, $G_{m,n}(z)$, $G_n(t)$ when the eccentricities λ and ν are 0.] Let $\hat{\sigma}^2$ be the usual unbiased estimate of σ^2 with $n-r$ degrees of freedom, and $\bar{X}_1,\ldots,\bar{X}_r$, the class means. We assert that $\sum f_j\xi_j > 0$, whenever $\sum f_j\bar{X}_j > \sqrt{(r-1)z}\ \sqrt{\sum f_j^2/n_j}\,\hat{\sigma}$, where $z = G^{-1}_{r-1,n-r}(1-\varepsilon)$, ε = the all over level of significance.

The clearance test is $\sum n_j(\bar{X}_j-\bar{X})^2 > (r-1)z\hat{\sigma}^2$. (We could have taken $(r-1)z = \Gamma^{-1}_{n-1}(1-\varepsilon)$ as in (1986), but we prefer the test which is exact in the normal case. We have $(r-1)G^{-1}_{r-1,n-r} \approx \Gamma^{-1}_{r-1}$ for $n-r$ large.)

The typical Scheffé situation is the case where we are prepared for surprises. We might e.g. discover, by looking at the data that the ξ depend upon a variable t, assuming values $t_1,\ldots,t_r$ associated with the r classes $j=1,2,\ldots,r$. To see how ξ increases and decreases with t, we consider effects $\xi_i-\xi_j$. To see if the increase is escalated (convex) we may consider effects $(\xi_i-\xi_j)/(t_i-t_j) - (\xi_k-\xi_\ell)/(t_k-t_\ell)$; $i>j\geq k>\ell$. Obviously in this situation Bonferroni could not be used.

Consider, however, the situation where a relation between the ξ and the t is suspected a priori. For simplicity, assume that the t_i are equidistant $t_i-t_{i-1} = 1$. We are interested in increases $\xi_i-\xi_j$ and curvatures $\xi_i-2\xi_j+\xi_k$.

According to Scheffé we would declare $\xi_i>\xi_j$ if

$$T = T_{diff} = (\bar{X}_i-\bar{X}_j)\Big/\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}\,\hat{\sigma} > \sqrt{(r-1)z} = C_S \quad \text{and} \quad \xi_i-2\xi_j+3\xi_k > 0 \quad \text{if}$$

$$T = T_{curv} = (\bar{X}_i-2\bar{X}_j+\bar{X}_k)\Big/\sqrt{\frac{1}{n_i}+\frac{4}{n_j}+\frac{1}{n_k}}\,\hat{\sigma} > \sqrt{(r-1)z} = C_S.$$

Using the Bonferroni method we have for each $\xi_i-\xi_j$; $i>j$; and each $\xi_i-2\xi_j+\xi_k$; $i>j>k$; a choice between three decisions (contrast > 0, contrast < 0, saying nothing). We have $M = \binom{r}{2}+\binom{r}{3}$ three-decision problems, each with level ε' (say). With all-over level ε, we then get by Bonferroni $\varepsilon' = \varepsilon/M$. Hence we state $\xi_i-\xi_j > 0$ or < 0 according as,

$$T_{diff} > G^{-1}_{n-r}\left(1-\frac{\varepsilon}{M}\right) = C_B \text{ or } < -C_B$$

and $\xi_i - 2\xi_j + \xi_k > 0$ or < 0 according as

$$T_{curv} > C_B \quad \text{or} \quad T_{curv} < C_B .$$

The T have Student distribution with $n-r$ degrees of freedom and eccentricity either

$$\delta = \delta_1 = (\xi_i - \xi_j)/\sigma\sqrt{\frac{1}{n_i} + \frac{1}{n_j}} = \Delta_1 / \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

or

$$\delta = \delta_2 = (\xi_i - 2\xi_j + \xi_k)/\sigma \sqrt{\frac{1}{n_i} + \frac{4}{n_j} + \frac{1}{n_k}} = \Delta_2 / \sqrt{\frac{1}{n_i} + \frac{4}{n_j} + \frac{1}{n_k}}$$

With $\varepsilon = 0.05$, $n = 30$, all $n_j = 6$, $r = 5$ we get

$$C_S = \sqrt{(r-1)z} = 3.322, \; M = 20, \; C_B = G_{n-r}^{-1}(1 - \frac{\varepsilon}{M}) =$$

$$= 3.075, \; \delta_1 = \sqrt{3}\Delta_1, \; \delta_2 = \Delta_2 .$$

We have introduced

$$\Delta_1 = (\xi_i - \xi_j)/\sigma \quad ; \quad \Delta_2 = (\xi_i - 2\xi_j + \xi_k)/\sigma .$$

We now get

Table 1. Probabilities of discovering true increases

Δ_1	By Bonferroni $T > 3.075$	By Scheffé $T > 3.322$
0	0.0025	0.001375
0.1	.0041	.0023
0.5	.0225	.0138
1.0	.1141	.0789
2.0	.6504	.5637
3.0	.9756	.9577
4.0	.9993	.9990

Thus if $(\xi_3 - \xi_1)/\sigma = 2$, then the probability of stating (correctly) that $\xi_3 > \xi_1$ is 65% in the case of Bonferroni and 56% in the case of Scheffé.

Table 2. Probabilities of discovering curvatures

Δ_2	By Bonferroni $T > 3.075$	By Scheffé $T > 3.322$
0	0.002500	0.001375
1	.0300	.0188
2	.1689	.1213
3	.4843	.3974
4	.8099	.7406
5	.9634	.9391
6	.9965	.9927

Thus if $(\xi_4-2\xi_2+\xi_1)/\sigma = -3$ the probability of (correctly) stating that $\xi_4 - 2\xi_2 + \xi_1 < 0$ is

$$G_{25}(-3.075, -3) = 1-G_{25} \quad (3.75, 3) = 48\%$$

by using Bonferroni, and $1 - G_{25} \quad (3.075, 3) = 40\%$ by using Scheffé.

The results are in accordance with Hjort's result that Bonferroni is often better if the statistical investigation is concerned with a finite number M of contrasts. However, the figures in the tables above tell us explicitely <u>how much better</u>. The differences are not impressive. It certainly is very pertinent to ask the following question. <u>How much are we willing to sacrifice in power (performance) with respect to special effects in order to allow for testing a priori unsuspected effects appearing when looking at the data</u>?

There can be no doubt, <u>the features of the performance function exposed by tables 1 and 2 give a beautiful insight into how the methods work and give a sound foundation for choosing the best method</u>.

It is the first line in the two tables corresponding to $\Delta_1 = \Delta_2 = 0$ which Hjort is concerned with. They give the "hypothetical" levels of the testing of a single inclination $\xi_i-\xi_j$ and curvature $\xi_i - 2\xi_j + \xi_k$, if it had been singled out a priori as the only effect to be tested. Of course they also define confidence intervals for the effects and their expected length, which is taken by Scheffé (1959) - and Hjort - as an indirect measure of performance (sensitivity). At the time when Scheffé wrote his monograph, the modern computer technologi was just emerging and his manner of comparing methods must be seen as a virtue of necessity. However, it is clear that the hypothetical level in the first lines of the tables are good indicators since lifting or lowering of them lifts or lowers the whole columns. The hypothetical levels will be equal if the number M of comparisons is such that $1 - \frac{0.05}{M} = 0.001357$, i.e. 36.34, i.e. M = 36. The two colunms in the table will be identical if M is equal to (the impossible) value 36.34.

4. The concept of performance function

In multiple comparisons an elementary decision d is a function from the space $\mathcal{F}$ of effects f to the two-point space (S,S^*), S meaning significance

$$d = \{d_f\}_{f \in \mathcal{F}}$$

The decision space $\mathcal{D}$ is the space of all d. A decision function ϕ is a function from the sample space of observations X to $\mathcal{D}$. The performance function of ϕ gives the probability

$$\beta_\phi(D|\eta)$$

of making a decision in D $(\subset \mathcal{D})$ when η is the true (unknown) parameter. For certain combinations of D and η it is desired that β should be large, for other combinations it should be small. Thus we study the goodness of ϕ by studying β. A numerical study of β is therefore important. A certain feature (i.e. a special D) has been studied in section 3 above. Modern computer facilities obviously make a more extensive study possible. Even the very sketchy study in section 3 indicates the importance of studying β. A more thourough study may involve extensive implimentation work. A suggestion of how to proceed is given in my paper (1986) p. 44, theorem 4, item (v), with the remark. The theorem deals with the general situation where a class $\mathscr{F}$ of contrasts relative to a null-state H_0 is considered. H_0 has w degrees of freedom. Consider a subclass $\mathscr{F}' \subset \mathscr{F}$, corresponding to a null-state $H_0' \supset H_0$ with $w' < w$ degrees of freedom. Let the clearance test with level ε have the form $Z > z$, where Z is the $-2 \log$ (likelihood ratio) or any other of the equivalent statistics. Then the probability of finding a significant contrast in $\mathscr{F}'$ is asymptotically equal to $\Pr(Z_{w'}(\kappa')>z)$, where $Z_{w'}(\kappa')$ has a chi-square distribution with w' degrees of freedom and eccentricity κ'. κ' is the "distance" of the true parameter from H_0' and is given by eq (40) in the paper. It assumes that distance is defined by a quadratic form with matrix equal to the information matrix. We shall refer to this as the $Z_w(\kappa)$-rule. Thus this gives the probability of finding a significant contrast, true or false. However, if κ' is large we might perhaps as a crude rule, assume that the probability of asserting a false contrast is small so that the above probability roughly gives the probability of finding true contrasts in $\mathscr{F}'$. But in any case we have an upper bound for the probability that we desire.

The example above gives us an opportunity of checking in special cases the conjecture that the upper bound may also be an approximation.

First consider the case when $\mathscr{F}'$ corresponds to the set of all contrasts for two means, i.e. the set of all $f_1\xi_1 + f_2\xi_2$, $f_1+f_2 = 0$ ($w'=1$), i.e. $f_1(\xi_1-\xi_2)$, i.e. really two contrasts, viz. $\xi_1-\xi_2$ and $\xi_2-\xi_1$. That is the "student" situation given above, where the tabled probabilities are the exact probabilities of discovering true contrasts. Compare now this with the probability of asserting contrasts

$$\Pr(|T|>|c|) = \Pr(T^2>c^2) = 1-G_{1,25}(c^2,\delta^2)$$

(Instead of the $Z_w(\kappa)$-rule we use the exact expression for probability of significance, since this is available in this case. The additional approximation is just approximation of the F-distribution by means of the chi-square distribution, which can be studied from standard tables.)

Table 3. Probability of significant $(\xi_1-\xi_2)$

Δ_1	Pr (false statement)	Pr (true statement)	Pr (significance)
0	.00375	.00375	.00751
0.1	$81\cdot 10^{-5}$	.00229	.00310
0.5	$8\cdot 10^{-5}$	.01383	.01391
1	$3\cdot 10^{-6}$	.07893	.07894
1.5	$4\cdot 10^{-7}$	.26605	.26606

(The three columns are computed independently)

It is seen that the approximation is fairly good for $\xi_i-\xi_j > \sigma/2$.

Now, let us consider the case when the subset of contrasts $\mathcal{F}'$ is the class of all $\xi_1 f_1+ \xi_2 f_2+ \xi_3 f_3$, $f_1+ f_2+ f_3 = 0$ $(w'= 2)$. The probability of asserting a significant contrast in $\mathcal{F}'$ is then given exactly by

$$\beta = 1-G_{2,25}(c^2/2,\kappa)$$

where $c^2/2 = (r-1)z/2 = 5.51784$

$$\kappa = \frac{6}{\sigma^2}\sum_{i=1}^{3}(\xi_i-\bar{\xi}_3)^2 = \frac{1}{\sigma^2}[3(\xi_1-\xi_3)^2+(\xi_1-2\xi_2+\xi_3)^2] = 3\Delta_1^2+ \Delta_2^2]$$

Hence if e.g. $\Delta_1 = 2$, $\Delta_2 = 1.5$, then the probability of significance equals 0.266018. [This is seen as follows. From the well known result

$$"\ \bigcup_f \left(\sum_1^r f_j\bar{X}_j > \sqrt{(r-1)z \sum_1^r f_j^2/n_j}\ \hat{\sigma}\right) \text{ if and only if}$$

$$\sum_{j=1}^r n_j(\bar{X}_j-\bar{X})^2 > (r-1)z\ \hat{\sigma}^2\ "$$

we get for $s < r$

$$\bigcup_f \left(\sum_{j=1}^s f_j\bar{X}_j > \sqrt{(s-1)\frac{r-1}{s-1}\, z \sum_1^s f_j^2/n_j}\ \hat{\sigma}\right)$$

if and only if

$$\sum_1^s n_j(\bar{X}_j-\bar{X}_{(s)})^2 > (s-1)\frac{r-1}{s-1}\, z\ \hat{\sigma}^2$$

Thus the probability of asserting a significant contrast in $\mathcal{F}'$ is equal to

$$1-G_{s-1,n-r}\left(\frac{(r-1)z}{s-1}, \kappa'\right)$$

$$\kappa' = \frac{1}{\sigma^2}\sum_{j=1}^s n_j(\xi_j-\bar{\xi}_{(s)})^2]$$

We know that the probability of asserting significant effects in $\mathcal{F}'$, all of which are significant, are less than β ; but we suggest that it is not much less. This conjecture is based on the fact that the probabilities of false assertions about $\xi_i-\xi_j$ and $\xi_i-2\xi_j+\xi_k$ are small according to table 3 above about $\xi_i-\xi_j$ and the corresponding table 4 about $\xi_i-2\xi_j+\xi_k$, viz.

Table 4. Probabilities of significant $(\xi_1-2\xi_2+\xi_3)$

Δ_2	Pr (false statement)	Pr (true statement)	Pr (significance)
0	.001375	.001375	.002750
0.1	$101\cdot10^{-5}$	.00185	.00286
1	$5\cdot10^{-5}$	.01877	.01882
1.5	$7\cdot10^{-6}$	.05222	.05223
2	$1\cdot10^{-7}$	.121133	.12133

Thus we might proceed as follows. Let A denote the assertion "at least one significant contrast in $\mathcal{F}$'". Let B be "no false statements" and hence B^* "at least one false statement". We know Pr(A) and want to evaluate $Pr(A\cap B)$. We have

$$Pr(A) \geq Pr(A\cap B) \geq Pr(A) - Pr(B^*)$$

Hence we need a lower bound for $Pr(A\cap B)$, which can be found by finding an upper bound for $Pr(B^*)$. We have

$$\frac{1}{\sigma}\sum_1^3 f_i\xi_i = (f_1+f_2/2)\Delta_1 - (f_2/2)\Delta_2$$

With $\Delta_1 = 2$ and $\Delta_2 = 1,5$ we make a false statement if and only if

(D_1) we assert $\Delta_2 < 0$,

or (D_2) we assert $\Delta_1+g\,\Delta_2 < 0$ for some $g > -\frac{4}{3}$,

or (D_3) we assert $\Delta_1+g\,\Delta_2 > 0$ for some $g < -\frac{4}{3}$.

In D_2 or D_3 there is obviously a strong dependency between different contrasts corresponding to different g. In addition each term in the unions D_2 and D_3 has probability $1 - G(3.322;\delta)$; $\delta = (\frac{4}{3} + g)(\frac{1}{3} + g)^{-\frac{1}{2}}$; which in the case of D_2 decreases from

.001375 to $3.8 \cdot 10^{-6}$ for $g = 1/4$ and then increases very slowly to $5.6 \cdot 10^{-6}$ for $g = \infty$. In the case of D_3 it increases from $5.6 \cdot 10^{-6}$ for $g = -\infty$ to .001375 for $g = -4/3$.

Hence it should be possible to evaluate the probabilities of the unions from a finite number - indeed a small number - of terms in the union close to $g = -4/3$. [One might even hope that terms corresponding to g far apart from $-4/3$ may be neglected and that terms (assertions) corresponding to g close to $-4/3$ are roughly equivalent to the term corresponding to $g = -4/3$. Hence $Pr(B^*) = 7 \cdot 10^{-5} + .001375 \times 2 = 0.00282$ and we make the rash statement that $0.26602 > Pr(A \cap B) > 0.26320$].

One way of proceeding would be as follows.

Introduce

$$T_g(X) = [\bar{X}_1 - \bar{X}_3 + g(\bar{X}_1 - 2\bar{X}_2 + \bar{X}_3)]/\hat{\sigma}\sqrt{\tfrac{1}{3} + g}$$

and for convenience

$$\sqrt{(r-1)z} = c, \quad a = -4/3 \quad \text{or more generally} \quad a = -\Delta_1/\Delta_2.$$

Hence

$$Pr(D_2) = Pr\{\bigcup_{g>a} (T_g(X) < -c)\}$$

We also introduce

$$R_{iN} = a + i \cdot 2^{-N}; \quad i = 1, 2, \ldots N \cdot 2^N; \quad N = 1, 2, \ldots \text{ ad.inf.}$$

Hence we have (rigorously),

$$Pr(D_2) = \lim_{N \to \infty} Pr[\bigcup_{i=1}^{N} (T_{R_{i,N}} < -c)]$$

[This is seen as follows. Introduce

$$S_g = \{X | T_g < -c\}, \quad S = \bigcup_{g>a} S_g, \quad S^N = \bigcup_{i=1}^{N} S_{R_{i,N}},$$

hence $Pr(D_2) = Pr(S)$, and

$$S^N \subset S^{N+1} \subset S \; ; \quad N = 1,2,\ldots$$

Obviously $\cup_{N=1}^{\infty} S^N \subset S$. But for any $x \in S$, there is a g such that $T_g(x) < -c$. From the density of the points $R_{i,N}$ and the continuity of $T_g(x)$ as a function g we can find an $R_{i,N}$ so close to g that we also have $T_{R_{i,N}}(x) < -c$. Hence $x \in S^N$ for some N and hence

$$\bigcup_{N=1}^{\infty} S^N = S$$

By the continuity of probability measures we then get

$$\lim_{N\to\infty} Pr(S^N) = Pr(S)$$

which is the same as the limiting result given above.]

It follows that we have

$$Pr(D_2) \approx Pr[\bigcup_{i=1}^{N} (T_{R_{i,N}} < -c)] = 1 - Pr[\bigcap_{i=1}^{N} (T_{R_{i,N}} > -c)]$$

For convenience we now write $T_{R_{i,N}} = T_i$, $R_{i,N} = R_i$. In order to compute this approximative probability, we note that we have

$$T_i = Y_i \sqrt{n-r}/\sqrt{Z}$$

where $(Y_i,\ldots,Y_N)$ is multinormal with means,

$$EY_i = (\Delta_1 + R_i\Delta_2)(\tfrac{1}{3} + R_i)^{-\frac{1}{2}} = \eta_i$$

and covariance matrix $\rho = (\rho_{ij})$,

$$\rho_{ij} = (\tfrac{1}{3} + R_iR_j)[(\tfrac{1}{3}+R_i^2)(\tfrac{1}{3}+R^2{}_j)]^{-\frac{1}{2}}$$

Z is independent of $(Y_1,\ldots,Y_N)$ and chi-square distributed with $n-r$ degrees of freedom. Thus we have

$$\Pr\left[\bigcup_{i=1}^{N}(T_i>c)\right] = \int\left[1-G_\rho(\eta_1-c\sqrt{v},\ldots,\eta_N-c\sqrt{v}\right)\gamma_{n-r}(v)dv$$

where G_ρ is the N-variate cumulative normal distribution means 0 and covariance matrix ρ.

Trial computations of this integral have to be made for ever increasing $N = 2,3,\ldots$ until it becomes "stationary". By the consideration and computations given above, the process might terminate after a few steps. However, not being a computer technolog, I am not able to judge the amount of work involved.

Concluding this section 4, I feel that Wald's statistical decision theory combined with modern computer technology, ought to create a revolution in judging statistical methods by studying the performance function. In the old days it was a great achievement (of R.A. Fisher) even to compute the critical points. When the concept of power was introduced, it was only of limited usefulness due to the computational difficulties. "Sample size needed" seems to be as far a one could go. Today these concepts ought to be depreciated, and even "uniformly best", "uniformly better than" must loose some of their importance unless they are quantified. (We should ask, "by how much?".)

5. The large sample approach

In the case of linear normal models Scheffé proved that his multiple comparison method has the following two properties.

(α) The method can be adjusted to a level of significance ε in the sense that the probability of obtaining at least one false contrast is at most equal to ε, regardless of the unknown value of the parameter η. Hence η may or may not belong to H_0.

(β) At least one estimated contrast surpasses the adjusted critical value if and only if the classical Fisher F surpasses its critical value.

Note that the result in (β) is a purely algebraic relation, it is neither probabilistic, nor asymptotic, nor approximate. The unknown parameter η does not occur in the algebraic relation and hence, trivially the property holds for all η.

The result in (β) is used to prove (α), but it has also an independent mission as a "clearance test" signaling that a contrast is present.

If we consider arbitrary densities $g(x;\eta)$ and contrasts that are arbitrary smooth functions $f(\eta)$ of η, it is shown in my (1986), theorem 4, that an adjusted delta method has properties similar to those in (α) and (β), but they are asymptotic and the analog to (β) is probabilistic.

The analog to (β) says inter alia that the probability that the clearance test is not significant whereas there is a significant contrast, goes to zero for any η, see theorem 4, (ii), (iv). Hence now η enters.

Thus the analog to (α) is concerned with the limit in distribution, whereas the analog to (β) is concerned with limit in probability, which justifies the use of the term clearance test.

Now, it is well known, and considered intuitively obvious from a statistical point of view, that the power of a reasonable test goes to 1 as the number of observations goes to infinity. (From a mathematical point of view it seems to be not so obvious, see my (1986) theorem 2(b), the proof p. 33-34 and lemma 4.) However, the intersection mentioned above between "clearance test not significant" and "a significant contrast" is obviously a subset of "clearance test not significant", which has probability = 1 - power $\to$ 0, if $\eta \notin H_0$.

Thus the statement in my paper about connection between clearance testing and contrast testing would be trivial if η were kept fixed not belonging to H_0. We have a similar situation in studying efficiency by non-parametric testing, the so-called Pitman-efficiency. If $\eta = \eta^{(n)}$ goes to H_0 as fast at $n^{-\frac{1}{2}}$ goes to 0, then good test can be distinguished from poor tests. By letting η be constant $\notin H_0$, any test, only marginally reasonable, is just as "good" as the Wilcoxen test!

The speed must of course be in "precipitate" direction, i.e. toward the foot point $\bar{\eta}^{(n)} \in H_0$ of $\eta^{(n)}$. (If e.g. H_0 were the perimeter of a circle, then the speed of $\eta^{(n)}$ should not be measured along a spiral surrounding and approaching H_0.) Thus a metric is needed. It appears that a convenient metric is a quadratic form with matrix equal to the information matrix. Then the power (probability of significance) attains a particularly nice form (see my (1986) eq. (31) and (41)). The foot point $\bar{\eta}^{(n)}$ is defined by the metric and $\sqrt{n}|\eta^{(n)}-\bar{\eta}^{(n)}| \to \Delta$ (say). (But other possibilities such as $\eta^{(n)} \to \eta \notin H_0$ are also included, they are just more trivial.)

These are the problems of clarification and proofs that I have taken up in my paper. It is possible that some of my proofs could be replaced by much simpler and more "transparent" proofs. At present I know of no such proofs. I don't think that the proofs of the fundamental property, given by me (1986) p. 45-46 and by Hjort (1988) p. 5-6 can be improved upon.

The limiting process $\eta^{(n)}$ also is of importance when discussing whether to use a priori estimates or null state estimates of η in the clearance tests and the variance of contrasts, discussed by Hjort section 5, remark A, see also my theorem 4 (iv), which justifies null state estimates. But, like Hjort, I feel that a priori estimated should be preferred. (However, there are delicate distinctions discussed in the litterature, see Eberhardt and Fligner, (1977)).

6. General and focalized parameters and contrasts

Concerning simplifications mentioned above, it might be said in general that simple proof can be obtained by considering simple propositions. In this connection I may also point to the distinction I make between "general formulations" and "reduced formulations" of the parameter space (1986) p. 23, and the different "quasi"- quadratic forms used for clearance testing, see eq. (38) and (39). ("quasi" because the matrix of the form depends upon the variables of the form.) The two test statistics are not obtained from each other by a transformation in the parameter space, I have taken the trouble of proving that they are equivalent up to limit in probability.

More important is the distinction I make between general contrasts, which may depend on the nuisance parameters in a certain manner, and "focalized contrasts" which depend upon the interest parameter only. That means that we may include nuisance parameters in the contrasts without increasing the degrees of freedom, and hence without impairing the power of the multiple comparison. An example of such a contrast is $\sum_1^w \nu_i f_i(\theta)$, where $\eta = (\nu, \theta)$, $\nu = (\nu_1, \ldots, \nu_w)$ is the interest parameter, θ nuisance, $\nu_1 = \ldots = \nu_w = 0$. H_0 and $f_1, \ldots, f_w$ are arbitrary smooth contrasts which generate $\mathcal{F}$.

7. The generality of the large sample multiple comparison method

Hjort's section 3 is interesting because he considers any situation with models implying asymptotically normal estimators (m.i.a. n.e.), and obtains methods which (i), do not depend upon "particulars of the statistical models" (ii), are not necessarily based on maximum likelihood estimators. Thus methods may be used in non-parametric situations, in autoregressive processed with Yule-Walker estimates etc., etc.

However, there are good reasons for paying special attention to models with independent groupwise identically distributed (i.g.i.d.) observations, using maximum likelihood estimators.

First, going through the proof of asymptotic normality and its consequences, and thinking in terms of approximations instead of limits, it seems likely that approximate normality must generally be good even for moderately large samples. (See my (1986) p. 30-33.) Also experiences show this to be the case.

Secondly, in the (i.g.i.d.) case the assumptions leading to asymptically distributed variables can be grossly simplified. They reduce essentially to assuming the existence of a solution of the maximum likelihood equations except on a set, the probability of which goes to zero, see my paper (1986) p. 38-40. Hence the verification of the basis for the asymptotic results is simple (but sometimes tricky).

Thus for i.g.i.d. observations the multiple comparison method is almost immediately available for application. In many other situations comprehensive numerical studies are needed.

8. Do there exist multiple comparison methods not "hampered" by null states? (Men nissen fulgte med på lasset, "But the puck went along on the van", from a Norwegian fairy tale)

Above I have commented upon the method in Hjort's section 3. In section 2 Hjort introduces a "general" method (by which is meant that the decision space is all-embracing). From the introduction it appears that he considers the method of section 2 to be not associated with a null state (hypothesis). This needs clarification. The rule of section 3 is at least as general as the rule of section 2. As a matter of fact it refers to a situation where the null state is such that the interesting parameters $\theta_1,\ldots,\theta_k$ (Hjort's notation) are completely specified, H_0: $\theta_1 = \theta_1^0, \ldots, \theta_k = \theta_k^0$. This applies whether the framework model is Hjort's m.i.a.n.e. or my i.g.i.d. (See my (1986) equation (4) p. 18, when $t = 0$, hence the ϕ_i constants.) Of course, such an hypothesis could be likelihood ratio tested by $-2 \log(\max L_{H_0} / \max L_{\text{a priori}})$, in case of maximum likelihood estimator, or in any case by a quasi-quadratic-form which follows from the asymptotically normal estimates that are chosen. With the special H_0, $f(\theta) = g(\theta) - g(\theta^0)$ is a smooth contrast for <u>any</u> smooth $g(\theta)$. By my theorem 4 (iii) about construction of confidence regions in the i.g.i.d. case or Hjort's argument in the m.i.a.n.e. case these confidence regions must have the form

$$\bigcap_f \left[|f(\theta^*) - f(\theta)| < K(X) \right]$$

where $f(\theta^*) = g(\theta^*) - g(\theta^0)$. But then we have $f(\theta^*) - f(\theta) = g(\theta^*) - g(\theta)$ and

$$\bigcap_g \left[|g(\theta^*) - g(\theta)| < K(X) \right]$$

where the intersection is taken over the smooth class of g. Thus Hjort's section 2 must really be superfluous.

Hjort's section 3 in the case of m.i.a.n.e. and my theory in the case of i.g.i.d. cover everything; more or less general contrast classes can be used, adjustable by means of null states. To a decreasing sequence of null states (sets of θ)

$$H_0' \supset H_0'' \supset \ldots \supset H_0$$

corresponds an increasing sequence of contrasts (sets of functions $f(\theta)$)

$$\mathcal{F}' \subset \mathcal{F}'' \subset \ldots \subset \mathcal{F}$$

Here H_0 is the one-point set $\{\theta^0\}$ and $\mathcal{F}$ is the class of all $g(\theta) - g(\theta^0)$, where g is smooth.

It is seen that in the confidence region for the g the θ^0 drops out. We obtain the same confidence region, regardless of the more specific choice of null state. But this is also true for any confidence region generated by an hypothesis $H_0^{(n)}$ (as in section 3 of Hjort's paper). This will be seen from the examples below.

It is stated above that the situation in section 3 is "at least as general" as the situation in section 2. But it is really not more general. The parameters may be transformed (see Hjort's remark after equation (13.4) and my (1986) p. 23-24)) from $(\theta_1,\ldots,\theta_k)$ to $(\nu_1,\ldots,\nu_k)$, where in the null state $\nu_1=\ldots=\nu_s=0$, whereas $\nu_{s+1},\ldots,\nu_k$ vary freely and may be pooled with the unmentioned other (nuisance) parameters and the non-parametricity.

It is sometimes more convenient to operate with a general parameter $\eta = (\eta_1,\ldots,\eta_v)$ and a null state restricting η to a t-space, see Sverdrup (1986) p. 18. Then we could add (and I ought to have added) the following "Remark" to theorem 4 (iii), p. 44.

"Consider the smooth class $\mathcal{G}$ of functions $g(\eta)$ which are constants $= c_g$, when η is in the null state. Then a simultaneous confidence interval for all $g \in \mathcal{G}$ is given by

$$\liminf \Pr\Big[\bigcap_{g\in\mathcal{G}}\big[\sqrt{n}(g(\eta^*)-g(\eta))<\sqrt{z}\ \sigma_g(\eta^*)\big]\Big] > 1-\varepsilon$$

This is seen from (86) by noting that $f(\eta) = g(\eta) - c_g$ is a contrast with estimate $f(\eta^*) = g(\eta^*) - c_g$ and standard deviation $\sigma_f = \sigma_g$. When $t = 0$ and hence $\eta = \eta^0$ (say) in the null state, then we have a confidence set for all smooth g".

Now to the examples.

Example 1. (Section 2 situation) Student's hypothesis with n observations and mean ξ. The null state is $\xi = 0$. Hence the set of possible effects $f\xi > 0$ reduces to two, corresponding to $f = 1$ or $f = -1$. We have a three-decision procedure giving $f = 1$ $(\xi>0)$,

$f = -1$ $(\xi<0)$ according as $\bar{X}$ is $> t_{1-\varepsilon}\hat{\sigma}/\sqrt{n}$ or $< -t_{1-\varepsilon}\hat{\sigma}/\sqrt{n}$ and, trivially, the Student test $|\bar{X}| > t_{1-\varepsilon}\hat{\sigma}/\sqrt{n}$ is the clearence test. To construct the confidence interval, we may just use the general rule given above, leading to $|\xi-\bar{X}| < t_{1-\varepsilon}\hat{\sigma}/\sqrt{n}$. We only have to consider the special null state $\xi = 0$, which completely (and mysteriously?) disappears in the confidence interval construction. We could have used null state $\xi = \xi_0$, with criterion $\bar{X} > \xi_0+t_{1-\varepsilon}\hat{\sigma}/\sqrt{n}$ etc. Then the contrasts would have been $f(\xi-\xi_0)$, $f = \pm 1$ with estimates $f(\bar{X}-\xi_0)$ and hence confidence interval $|\xi-\xi_0-(\bar{X}-\xi_0)| < t_{1-\varepsilon}\hat{\sigma}/\sqrt{n}$. Hence ξ_0 cancels out!

Example 2. (Section 3 situation). Consider the one-way lay out in analysis of variance with expectations $\xi_1,\ldots,\xi_r$. We want a simultaneous confidence interval for the $g(\xi) = \Sigma f_j\xi_j$, $\Sigma f_j = 0$, hence $g(\xi) = \Sigma_1^{r-1} f_j(\xi_j-\xi_r)$. Let us again do this by introducing the usual null state $\xi_1 =\ldots= \xi_r$. Then the $g(\xi) = f(\xi)$ are contrasts and we get the significance criterion

$$\sum f_j\bar{X}_j > \sqrt{(r-1)c\Sigma f_j^2/n_j}\,\hat{\sigma} = K_f(X),$$

where c is the Fisher-fractile. The clearance test is the ordinary F-test $\Sigma\, n_j(\bar{X}_j-\bar{X})^2 > (r-1)c\hat{\sigma}^2$. The confidence region is now

$$\bigcap_f (|g(\xi) - \Sigma f_j\bar{X}_j| < K_f(X))$$

Consider now another null-state, viz.

$$\xi_j = \xi_r + \Delta_j;\ j=1,2,\ldots,r-1,$$

(Δ_j known, define $\Delta_r = 0$). Then the contrasts are

$$f(\xi) = \sum f_j\xi_j - \sum f_j\Delta_j; \qquad \sum f_j = 0$$

Of course we have another significance criterion

$$\sum f_j\bar{X}_j > \sum f_j\Delta_j + K_f(X)$$

and another clearance test

$$\sum n_j(\bar{X}_j-\bar{X}-\Delta_j+\bar{\Delta})^2 > (r-1)c\hat{\sigma}^2$$

But since

$$g(\xi^*) = \sum f_j\bar{X}_j - \sum f_j\Delta_j$$

we have

$$g(\xi)-g(\xi^*) = \sum f_j\xi_j - \sum f_j\bar{X}_j$$

and hence the same confidence region as before. Thus again, in constructing the confidence interval, we are not hampered by the special choice of null state.

Example 3. (Section 2 situation) The results of tossing n times with a die have been observed. We want to say as much as possible about how the probabilities $\pi_1,\ldots,\pi_6$ compare with each other and with the probabilities $\frac{1}{6},\ldots,\frac{1}{6}$ for a fair die.

Taking the situation a little more general, let $\pi_1,\ldots,\pi_r$ be the probabilities of r possible outcomes in one trial and $X_1,\ldots,X_r$ the frequencies of these outcomes in n independent trials. We use the notations $\pi = (\pi_1,\ldots,\pi_r)$, $X = (X_1,\ldots,X_r)$.

According to what we have said above, we consider a smooth class of functions $g(\pi)$ containing linear and log-linear subclasses (see below). Let $\pi = \pi^0$ be the null state. Then any $f(\pi) = g(\pi)-g(\pi^0)$ is a contrast in $\pi_1,\ldots,\pi_{r-1}$ and we assert that $f(\pi) > 0$, if $f(\pi^*) > \sqrt{z/n}\,\hat{\sigma}_f$ where $z = \Gamma_{r-1}^{-1}(1-\varepsilon)$ and

$$\sigma_f^{*2} = \text{as.var}\sqrt{n}\, f(\pi^*) = \sum f_j^2(\pi^*)\pi_j^* - (\sum f_j(\pi^*)\pi_j^*)^2,$$

$$\pi_j^* = X_j/n, \quad f_j(\pi) = \partial f/\partial\pi_j = \partial g/\partial\pi_j = g_j(\pi)$$

(see my (1986) p. 54 eq.(124)). We may use as clearance test statistic either

$$Z_1 = -2n \log n - 2 \sum X_j \log(\pi_j^0/X_j)$$

or

$$Z = \sum_j (X_j - n\pi_j^0)^2/X_j$$

Thus $Z_1 > z$ (resp. $Z>z$) indicates that contrasts are present.

Special subclasses of contrasts are the linear, where

$$f(\pi) = \sum c_j(\pi_j - \pi_j^0), \quad \text{var}\sqrt{n}\, f(\pi^*) = \sum c_j^2 \pi_j - (\sum c_j \pi_j)^2$$

and the log-linear

$$f(\pi) = \sum c_j \log \pi_j/\pi_j^0, \quad \text{as.var}\sqrt{n}\, f(\pi^*) = \sum c_j^2/\pi_j - (\sum c_j)^2$$

[We may note in passing that by my (1986) theorem 7, or my (1975)), for any n, a linear contrast is present if and only if $Z > z$. Thus we have an exact algebraic relationship, we need not resort to "limit in probability". See also Goodman (1964) which treats a similar situation.] We now again apply my theorem 4, (iii) equation (86), see also theorem 6, equation 126), to obtain the simultaneous confidence intervals

$$A: \quad \bigcap_c \left[\sqrt{n}\left|\sum c_j \pi_j^* - \sum c_j \pi_j\right| < \sqrt{z\sum(c_j - \bar{c})^2 \pi_j^*}\right]$$

$$B: \quad \bigcap_c \left[\sqrt{n}\left|\sum c_j \log\pi_j^* - \sum c_j \log\pi_j\right| < \sqrt{z(\sum c_j^2/\pi_j^* - (\sum c_j)^2)}\right]$$

Here $\bar{c} = \Sigma c_j \pi_j$. We then have (inter al.)

$$\liminf_{n\to\infty} \Pr(A \cup B) \geq 1 - \varepsilon$$

since both A and B have as subset

$$C: \quad \bigcap_g \left[\sqrt{n}\left|g(\pi^*) - g(\pi)\right| \leq \sqrt{z(\sum g_j(\pi^*) - (g_j(\pi^*)\pi_j^*)^2)}\right]$$

and $\liminf \Pr(C) \geq 1-\varepsilon$. g is taken over a class of g that is smooth. (This requires the $|c_i|$ in the definition of the set B to be bounded.) The sets A, B, C, may be taken as sets of trials (tosses) or as sets of $(X_1,\ldots,X_n) = X$.

We would have obtained the same result if, in place of the arbitrary choice π_j^0 we had made the specific choice $\frac{1}{r}$ ($= \frac{1}{6}$ for the die). This is similar to example 2, where the arbitrary choice of Δ_j gives the same confidence region as with the specific choice of all $\Delta_j = 0$, and to example 1, where it suffices to define the null state as $\xi = 0$, not an arbitrary $\xi = \xi_0$. [This is contrary to J. Neyman (1937) who from (optimum) tests of $H_0 : \eta = \eta^0$ generates (optimum) confidence regions for η by varying the arbitrary η^0. See also Lehman (1986) p. 213-216.]

Thus it is seen that we may use many different specific null states and obtain the same confidence region from the tests as long as all the null states have the same degree of freedom, which determines z. This may seem rather curious.

However, what is more important is that all the situations which Hjort and I consider are all "hampered" by (or rather "blessed" by) a null state, useful in the case of multiple testing. Above all, the clearance testing throws light on the classical tests, Fisher's F-test and the chi-square test (see my (1986) p. 15).

9. Why multiple comparison tests? Why not "just" simultaneous confidence intervals?

This questions often pops up in discussions of multiple comparison problems, and in many other context in statistics. It is therefore natural to take it up here (even if it is not commented upon in Hjort's paper).

In fact, simultaneous confidence intervals are seldom used in practical statistical work. However, they are often used to describe a test procedure for multiple comparison. As such it is a misnomer, compared to the original meaning of confidence intervals, which were meant to be terminal decisions. Thus in the case of the one-way lay out in analysis of variance, considered above (in chapter 3), with 5 classes, 6 observations in each class, all over level 0.05, observed class means 375, 470, 367, 296, 363, and estimated variance $\hat{\sigma}^2 = 1662.5$; we have as simultaneous confidence intervals for all possible effects $\Sigma_1^5 f_i\xi_i$; $(\Sigma f_i=0)$,

$$\bigcap_f \{|\sum_1^5 f_i\xi_i - 375f_1 - 470f_2 - 367f_3 - 296f_4 - 363f_5| \leq 55.3\sqrt{\sum_1^5 f_i^2}\}$$

As a terminal decision this would usually say nothing.

However, let us test which interesting contrasts may be present. For contrasts of the form $\xi_i-\xi_j$ we find $\bar{X}_i-\bar{X}_j > 55.3\sqrt{1^2+1^2} = 78.2$ from which we find that treatment 2 may be declared better than the other treatments and treatment 4 inferior to treatment 1. We find no other significant $\xi_i-\xi_j$.

Suppose that a "t-property" of the treatments have been measured as

$$t_1 = 16.9, \quad t_2 = 23.3, \quad t_3 = 14.9, \quad t_4 = 10.3, \quad t_5 = 15.3$$

respectively for the 5 classes. Arranging them according to ascending values of t, we get

Treatment no.	4	3	5	1	2
t	10.3	14.9	15.3	16.9	23.3
$\bar{X}$	296.3	366.8	363.3	374.8	470.2
Slopes		15.3	- 8.8	7.2	14.9

$(15.3 = (366.8-296.3)/(14.9-10.3)$, etc.)

We might believe that we have "discovered" a dependence of population mean ξ on t, not suspected in advance. We might take

$\beta = \Sigma(\xi_i(t_i-\bar{t})$ as a measure of all over decrease or increase with t (without assuming linearity). The estimate is $\beta^* = \Sigma\ \bar{X}(t_i-\bar{t}) =$ 1161 and the critical point is $55.3\sqrt{\Sigma(t_i-\bar{t})^2} = 519$. Thus we assert dependence. Is there any local escalation (curvature)? Is e.g.

$$7.2 = (\bar{X}_2-\bar{X}_1)/(t_2-t_1) - (\bar{X}_1-\bar{X}_5)/(t_1-t_5)$$

significant? Since $t_2-t_1 = 6.4$ and $t_1-t_5 = 1.6$ the above expression for the curvature may be written

$$-\ 0.781\ \bar{X}_1 + 0.156\ \bar{X}_2 + 0.625\ \bar{X}_5$$

Hence the critical point for the observed curvature is equal to $55.3\sqrt{0.781^2+0.156^2+0.625^2} = 56.0$. Thus we can not claim escalated effect of t on ξ.

Now, the manner of carrying out the multiple testing could of course be read out of simultaneous confidence intervals for $\Sigma\ f_i\xi_i$ given above. Thus it seems that the whole <u>question of confidence intervals versus multiple testing is just formalism. But it is definitely not</u>!

There are situations where simultaneous confidence intervals are natural as terminal decisions. But these are different from situations where the purpose is to find interesting effects. In the first case we may want to have the <u>confidence intervals narrow</u> in the metric sense. In the second case this is not interesting. All we then want is <u>high probability of discovering true effects</u>. This means that the <u>performance function is judged differently in the two cases</u>.

Consider the Student situation with two samples of size m and n, population means ξ and η, variance σ^2, and effects $f_1\xi_1 + f_2\xi_2$; $f_1+f_2 = 0$. Hence we only have to consider $\pm(\xi_1-\xi_2)$. The $1-\varepsilon$ confidence interial is

$$\xi_1-\xi_2 \in \bar{X}-\bar{Y} \pm c\sqrt{\frac{1}{m}+\frac{1}{n}}\,\hat{\sigma}\,, \qquad c = G^{-1}_{m+n-2}\,(1-\varepsilon)$$

where $\bar{X}$, $\bar{Y}$, $\hat{\sigma}$ are the usual estimates of ξ, η, σ. The length of the interval is

$$L = 2c\sqrt{\frac{1}{m}+\frac{1}{n}}\,\hat{\sigma}$$

the distribution of which is derived from Γ_{m+n-2}. Thus the central Student and chi-square distributions G_{m+n-2} and Γ_{m+n-2} are involved.

In the case of a "multiple" comparison, assert that $\xi_1-\xi_2 > 0$ or < 0 according as

$$\bar{X}-\bar{Y} > \text{ or } < c\sqrt{\frac{1}{m}+\frac{1}{n}}\,\hat{\sigma}$$

the probabilities of which are respecively $1-\beta$ and β, where

$$\beta = G_{m+n-2}(c;(\xi_1-\xi_2)/\sigma\sqrt{\frac{1}{m}+\frac{1}{n}})$$

Thus the essential features of the performance function is described by the eccentric Student distribution $G_{m+n-2}(\cdot;\kappa)$.

Now, one might object that why worry? "Any" good test gives rise to a good confidence interval. However, this kind of reasoning may be carried further. "Any" good point estimator gives rise to a good test and a good confidence interval. Hence the only thing that matters is the theory of point estimation.

I think that the tables 1 and 2 with comments in chapter 3 above should show convincingly the importance of studying the probabilities of the different decisions which the statistical investigation is aiming at. This will enable us to confront possible states of nature (viz, a parameter θ) with possible decisions d. For this purpose one might, following Wald (1950), introduce a weight function (loss function). This may not be deemed necessary or convenient. But the principal idea embodied in the loss function cannot be neglected.

Thus from a methodical point of view there is a real difference between a situation where a priori considerations lead us to aim toward a set of assertions about effects, and a situation where a priori considerations lead us to aim toward a simultaneous confidence intervals as terminal decision.

On the other hand it may happen that the outcome of looking at the data is inter al. that simultaneous confidence intervals is desirable. To see this, write η on a reduced form $(\nu_1,\ldots,\nu_w,\theta_1,\ldots,\theta_{v-w})$, where the θ_i vary freely, $\nu_1=\ldots=\nu_w=0$ is the null state and all effects are focalized, i.e. of the form $f(\nu)$. Then, looking at the observations we might take interest in a set of smooth functions $g_t(\nu)$; $a<t<b$; t varying continuously. Since then all $\pm f_t(\nu) = \pm[g_t(\nu)-g_t(0)]$ are effects with estimates $f_t(\nu^*) = g_t(\nu^*)-g_t(0)$, $f_t(\nu^*)-f_t(\nu) = g_t(\nu^*)-g_t(\nu)$, then we have from (86) in my (1986) the confidence band

$$\bigcap_t \left[\sqrt{n}\,|g_t(\nu^*)-g_t(\nu)|<\sqrt{z}\ \sigma^*_{f_t}\right]$$

which has probability at least $1-\varepsilon$ in the limit $(z = \Gamma_w^{-1}(1-\varepsilon))$. An example of such a confidence band for a linear regression is given in my (1976) paper.

References

Eberhardt, K.R. and Fligner, M.A.(1977). A comparison of two tests for equality of two proportions. The American Statistian, 31, p. 151-7.

Goodman, L.A.(1964). Multinomial confidence interval for contrasts among multinomial populations. Am.Math.Stat. 35.

Hjort, N.L.(1988). On large-sample multiple comparison methods. Scand.J.Stat. 15.

Lehman, E.L.(1986). Testing Statistical Hypotheses. John Wiley & Sons. New York.

Neyman, J.(1937). Outline of a theory of statistical estimation based on the classical theory of probability. Phil.Trans.A. 236.

Scheffé, Henry (1959). The Analysis of Variance. John Wiley & Sons. New York.

Sverdrup, E.(1976). Significance testing in multiple statistical inference. Scand.J.Stat. 3.

Sverdrup, E.(1977). The logic of statistical inference: Significance testing and decision theory. Bull.Int.Stat.Inst. XLVII (1).

Sverdrup, E.(1986). Multiple comparison and the likelihood ratio testing: General theory and application to categorical data. Scand.Actuarial J. 13-63.

Corrections to my paper "Multiple Comparison and the Likelihood Ratio Testing". Scand. Actuarial J. 1986: 13-63.

By Erling Sverdrup

The correction *** corrects a misleading error. The "corrections" * imply leaving out a superfluous assumption. Corrections ** add an assumption needed.

	p. 14	line 22 ↓	:	"ideals" → "ideas"
	p. 27	" 11 ↑	:	"$(\sqrt{n}$" → "$\sqrt{n}($"
	p. 28	" 13 ↓	:	"η_j" → "η_{w+j}"
	p. 28	" 18 ↑	:	"$\phi(\bar{\eta})^{-1}$" → "$\sigma(\bar{\eta})^{-1}$"
*	p. 29	" 12-11 ↑	:	"The test of H_0 based on Z_0' is" → "The tests of H_0 are"
*	p. 34	" 13-12 ↑	:	Interpolate between the two lines "let $\varepsilon_n = 1$ and hence $\nu^{(n)} \to \Delta^I$. We still have plim $Z = \infty$. (We also have "
	p. 34	" 11 ↑	:	"$= \infty$" → "$= \infty)$"
**	p. 40	" 7 ↓	:	Add to the sentence ", if η^* exists (eq.(64))"
**	p. 40	" 1 ↑	:	"Then the " → " If η^* and $\hat{\eta}$ exist (eq.(64) and (66)), then"
*	p. 43	" 13-12 ↑	:	Leave out: "See also lemma 4 ... below"
*	p. 44	" 1-2 ↓	:	Leave out: "If the likelihood ... then the"
*	p. 44	" 13-14 ↓	:	Leave out: "If it is assumed ... consistent"
*	p. 45	" 10-15	:	Replace everywhere "ε" by "a" (In order not to confuse with level of significance)
	p. 47	" 4 ↑	:	"$\tilde{f}$" → "f"
	p. 50	" 6 ↓	:	Leave out "η ="
***	p. 50	" 10-21 ↓	:	↓: Replace everywhere subscript "12" by "11"
	p. 50	" 14 ↓	:	Leave out "$(\bar{X}, \eta)$"

Oslo, April 1988